Invited Paper

Image processing applications of optical neural networks

Demetri Psaltis, Hsin-Yu Sidney Li, and Xin An

California Institute of Technology
Pasadena, California 91125

1. INTRODUCTION

The basic architecture for a holographic neural network [1] consists of 2-D neural planes that are implemented using some form of spatial light modulator (SLM) or "smart pixel" technology and 3-D holograms that implement the weighted interconnections between the neurons or pixels on the SLMs. The 3-D storage capability of the volume hologram provides the system with very high storage density allowing us to implement efficiently very large networks. Typically, a single volume hologram can store $10^9 - 10^{10}$ weights in a volume smaller than 1 cm^3. Therefore, optical networks with billions of weights can be readily realized [2]. In addition to the large size of the optical networks, speed is another desirable property. High speed is derived from the 2-D parallelism of the optical system and the SLM in particular. Currently, the number of pixels on commercially available SLMs is approximately 500×500 and devices with one million pixels are expected to become available in the near future with the advent of high definition television. The speed of the network is also dependent on the switching speed of the SLM. The devices most commonly used today are based on nematic liquid crystal technology and they typically have a switching speed of approximately 30 milliseconds (video rates). Devices based on ferroelectric liquid crystals or semiconductors (GaAs) are expected to provide switching speeds in the 100 microsecond regime and possibly less. We can get an estimate for the processing speed of a holographic network by dividing the total number of weights stored in the hologram by the switching speed of the SLM. For current technology the processing speed is in excess of $10^9/(3 \times 10^{-2}) \approx 3 \times 10^{10}$ weight updates per second. Increasing the SLM speed to a millisecond yields processing speed in excess of 10^{12} weight updates per second. The performance of optical networks (a billion adaptable weights and 10^{12} weights updates per second) cannot be easily matched by electronic implementations. However, some applications may require even better performance. Specifically, networks that are trained with local algorithms [3,4] in image recognition applications, can have a huge storage requirement. The use of optical storage can provide an effective solution to this large storage requirement [5]. In this paper we explore the use of 3-D disks [6] for the construction of networks with extremely large storage capacity. 3-D disks can store up to 10^{12} weights [7] per disk. In this paper we discuss how 3-D disks are used to implement an optical neural network and then derive the capacity and speed of the resulting architecture.

2. 3-D DISK NETWORK

The optical network implemented with a 3-D disk is shown in Figure 1. A 3-D disk is simply a holographic material (e.g., photorefractive crystal or photopolymer) shaped as a disk. At each location on the disk information is stored in the volume of the material through angular or

0-8194-1275-9/93/$6.00

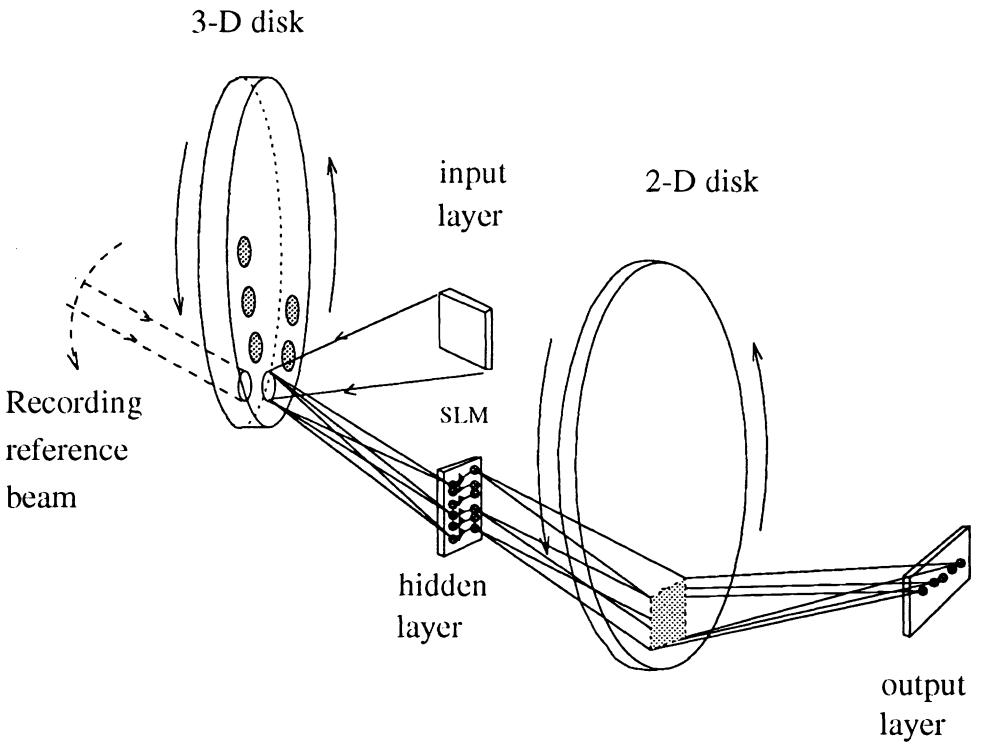

Figure 1. Optical Network implemented with a 3-D disk.

wavelength multiplexing as in conventional holographic storage. The typical area of each location is several mm^2. The transverse area of the disk is large enough to support many such locations. Therefore information is both spatially and angularly multiplexed in a 3-D disk. Different locations on the disk are accessed by using the mechanical motion of the disk and the head. The disk's rotation allows access to different spots around the disk whereas radial motion of the readout head provides the scanning along the second direction. The holographic network shown in Figure 1 uses the holograms stored at each location in the conventional manner of a holographic interconnected network [1]. Then different locations are used to extend the network's capacity. There are two distinct modes of operation. In the first mode, each location is used as an independent network that is separately trained. Then the different locations on the disk are used to store the weights for different tasks. For instance, if the network is trained to recognize faces, then each location can be devoted to store the weights for a network that recognizes a specific person. Multiple persons can be searched for by rotating the disks so that different areas are illuminated by the optical beam, effectively reprogramming the network to look for a different person each time a new location is illuminated. In the second mode of operation multiple locations on the disk are used to build-up a network that is larger than the capacity of a single location. This is done by first evaluating the network at one location and storing the response in an electronic memory, prior to thresholding at the output layer. The response of subsequent layers is then accumulated in the electronic memory thereby constructing a large "virtual" hidden layer that is built-up over time. The final thresholding is performed electronically after all the locations that make up the network have been evaluated.

The network described above can either implement a very large network with the number of

weights being limited by the overall capacity of the disk or a large number of smaller networks (each with approximately a billion weights) that can be quickly reprogrammed by rotating the disk. In what follows we will consider the factors limiting the performance of such systems and derive their storage capacity (the maximum number of weights that can be supported) and their processing speed (the number of synapses per second that the network can implement).

3. NETWORK SIZE

The total number of weights, N_w, that the system in Figure 1 can support is given by

$$N_w = N_i N_h N_l + N_h N_o N_l \tag{1}$$

where N_i is the number of input pixels, N_h is the number of hidden units that can be implemented at one time (at a single location), N_o is the number of output units, and N_l is the number of locations on the disk where holograms are recorded. In image processing applications it is often the case that $N_i >> N_h >> N_0$ which implies that $N_w \approx N_i N_h N_l$. The input SLM sets the limit on N_o to approximately one million pixels. N_h and N_l are limited by the properties of the 3-D disk. We wish to maximize the number of locations on the disk and at the same time maximize the number of hidden units. Unfortunately, the two are related. In order to increase the number of hidden units, we must record holograms at each locations using reference beams with a larger angular deviation. The increase in the angle of the reference beam, causes a larger area on the disk to be illuminated. This effect is shown in Figure 2. Therefore there is a maximum number of hidden units beyond which the increase in the illuminated area is not compensated by the increase in the number of hidden units. This effect determines the optimum number of hidden units to be used.

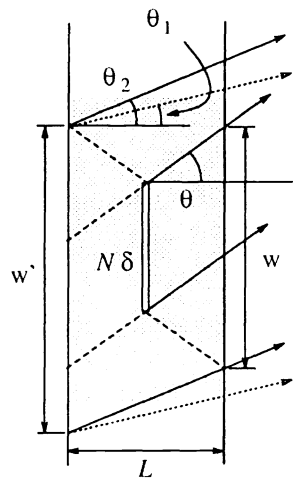

Figure 2. Extra area taken up by defocusing and reference beam angle change.

Another parameter that needs to be optimized is the thickness of the disk. The Fourier transform of the input image is shown to come to focus at one plane inside the crystal, halfway

between the two disk surfaces. The extend of the signal beam is smallest at the Fourier transform plane and it expands away from it. Therefore the area illuminated on the surface of the disk is larger than the spatial extend of the Fourier transform itself and it grows larger as the disk becomes thicker. This effect determines the optimum disk thickness. The two competing mechanisms that give rise to the tradeoff are as follows: As the disk becomes thicker, the angular selectivity increases and therefore we can have more hidden units within the same angular range of the reference beams. However, the increase in the thickness also causes the illuminated area to grow because of the defocusing just described. It is possible to optimize all the parameters of the system simultaneously [7] to maximize the storage capacity of the disk. A convenient measure for expressing the storage density is in terms of the number of weights per unit area on the disk's surface. This allows us to predict the storage density independently of the size of the disk and also provides a direct comparison with the density of storage that can be obtained with a network implemented with 2-D disks [8]. The optimum storage is plotted in Figure 3 where the number of weights per μm^2 that can be stored on the disk is plotted as a function of the disk's thickness. Notice that the density peaks at 47.7 weights per micron squared at a thickness of approximately 3.2 mm. Also plotted on the same figure is the corresponding optimum number of hidden units that achieve the optimum density. For the maximum density, the number of hidden units per location is $N_h = 250$. The area that is illuminated on the disk in this optimum configuration is 5.27 mm^2. Therefore, if the diameter of the disk is 5.25" the number of locations is $N_l = 2411$ (assuming a 2 cm radius area at the center of the disk that is not used). Substituting $N_i = 10^6$, $N_h = 250$, and $N_l = 2650$ in Eq. (1) we obtain a network size of 6.03×10^{11} weights.

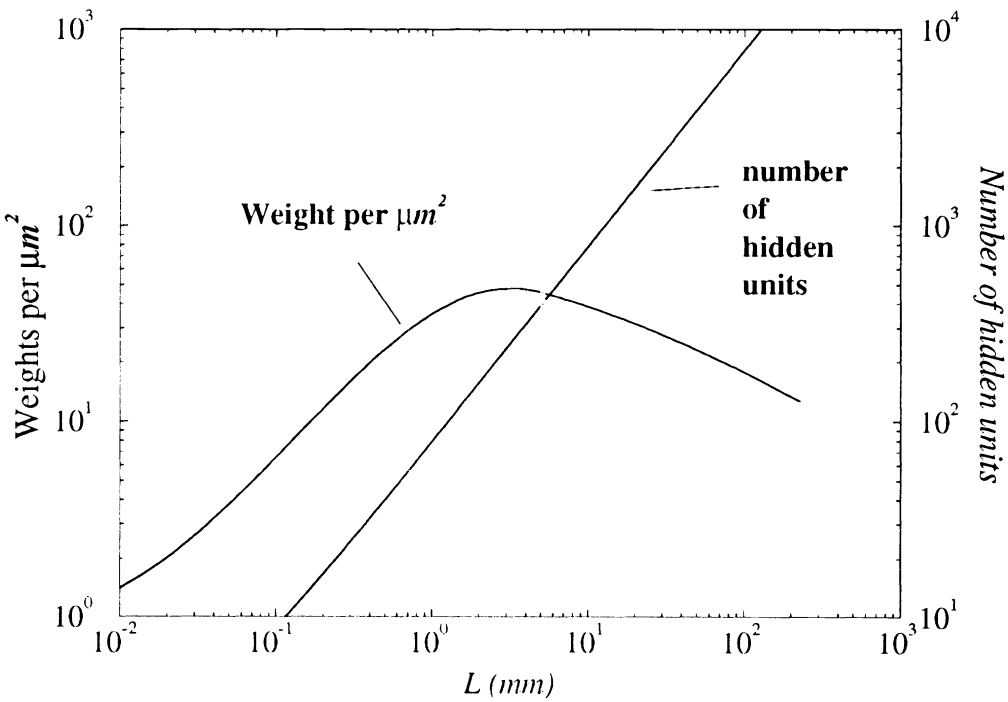

Figure 3. Optimum storage density as a function of crystal thickness.

4. PROCESSING SPEED

The processing speed of the network in Figure 1 (measured in weights per second) is equal to the number of weights per location divided by the time τ it takes to evaluate the response of the network at each location. We have already calculated the number of weights per location to be $N_i N_h = 2.5 \times 10^8$. The time τ is determined by the sensitivity of the hidden units. If we are given the available optical power and the light efficiency of the system, then the sensitivity of the devices used to implement the hidden units determine the required integration time. Specifically

$$\tau = \frac{MhcN_h}{\eta_{SLM}\eta_{HOL}\eta_{DET}I\lambda}, \tag{2}$$

where M is the number of photons that each hidden unit requires to turn on, h is Planck's constant, c is the speed of light, η_{SLM}, η_{HOL}, and η_{DET} are the SLM, detector, and hologram efficiencies, I is the incident light intensity, and λ is the wavelength of light. Using $M = 10^3$, $N_h = 250$, $I = 100\ mW$, λ =488 nm, $\eta_{HOL} = 10^{-5}$, $\eta_{DET} = .5$, and $\eta_{SLM} = .1$, $\tau = 2.037\ \mu sec.$ The processing speed of the system is then equal to 1.23×10^{14} weights per second. The rotational speed of the disk required to sustain this processing speed is about 3,600 revolutions per minute [9] which can be easily achieved.

5. ACKNOWLEDGMENT

The support of DARPA and AFOSR are greatly appreciated.

6. REFERENCES

[1] D. Psaltis, D. Brady, and K. Wagner, "Adaptive Optical Networks Using Photorefractive Crystals," *Appl. Opt.*, vol. 27, no. 9, pp. 1752–1759, 1988.

[2] H.-Y. S. Li, Y. Qiao, and D. Psaltis, "Optical Network for Real-Time Face Recognition", *Appl. Opt.*, to appear Sept. 1993.

[3] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons, New York, 1973.

[4] T. Poggio, and F. Girosi, *A Theory of Networks for Approximation and Learning*, MIT AI Laboratory and Center for Biological Information Processing, Whitaker College, AI Memo 1140, CBIP paper 31, 1989.

[5] M. A. Neifeld and D. Psaltis, "Optical Implementation of Radial Basis Classifiers," *Appl. Opt.*, vol. 32, no. 8, pp. 1370–1379, 1993.

[6] D. Psaltis,"Parallel optical memories", *Byte*, vol. 17, no. 9, pp. 179–182, 1992.

[7] H.-Y. Li and D. Psaltis, "3-D Optical Disks", submitted to *Applied Optics*.

[8] Demetri Psaltis, Mark Neifeld, Alan Yamamura, and Seiji Kobayashi, "Optical Memory disks in Optical Information Processing", *Appl. Opt.*, vol. 29. no. 14, pp. 2038–2057, 1990.

[9] H.-Y. Li, *Photorefractive 3-D disks for optical data storage and artificial neural networks*, PhD Thesis, California Institute of Technology, 1993.